On Generative Spoken Language Modeling from Raw Audio

Year: 2021 | Citations: 421 | Authors: Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak

---

Abstract

Abstract We introduce Generative Spoken Language Modeling, the task of learning the acoustic and linguistic characteristics of a language from raw audio (no text, no labels), and a set of metrics to automatically evaluate the learned representations at acoustic and linguistic levels for both encoding and generation. We set up baseline systems consisting of a discrete speech encoder (returning pseudo-text units), a generative language model (trained on pseudo- text), and a speech decoder (generating a waveform from pseudo-text) all trained without supervision and validate the proposed metrics with human evaluation. Across 3 speech encoders (CPC, wav2vec 2.0, HuBERT), we find that the number of discrete units (50, 100, or 200) matters in a task-dependent and encoder-dependent way, and that some combinations approach text-based systems.1